A Review of the DC IMPACT Teacher Evaluation System

Drew H. Gitomer, Kevin Crouse, and Jeanette Joyce
Graduate School of Education, Rutgers University

Introduction

With the design and implementation of the DC IMPACT evaluation model, Washington DC has been at the forefront of educator evaluation initiatives that states are developing across the country. The U.S. Department of Education has encouraged these initiatives primarily through two mechanisms. Initially, the U.S. Department of Education provided funding through the Race to the Top program for states to voluntarily implement new evaluations that included student achievement as a substantial focus, and it more recently began requiring states and districts to implement similar evaluations to waive out of certain Elementary and Secondary Education Act (ESEA) requirements or to continue to qualify for existing programs such as the School Improvement Grants and the Teacher Incentive Fund. This report provides an overview of IMPACT, Washington DC's evaluation and performance management system, which was initiated in 2009 and aimed at improving educator effectiveness. However, this report focuses on classroom teachers and not on all of the DC educators and school staff who fall under the umbrella of IMPACT.

In our studies of teacher evaluation systems and as the authors of this report, we describe modern teacher evaluation systems by asking critical questions to understand how they are oriented. First, what decisions and actions is the system intended to support? In the case of IMPACT, the system is intended to provide a basis for feedback and support to teachers, differential recognition and compensation, and disciplinary action.

Second, how is the system designed and implemented? What are the constituent measures that contribute to the evaluation and what features of measurement design and implementation contribute to or detract from valid and reliable inferences about teacher quality? IMPACT

aggregates measures of student growth, classroom practice, practice outside the classroom, and professionalism into a composite evaluation score that is used to make consequential decisions about teachers.

Third, what decisions are made and actions taken on the basis of the evaluation and what is known about the impact of those decisions and actions? IMPACT includes a set of explicit decisions that affect the employment and compensation status of teachers. Some studies describe the distribution of district teachers on both the component and aggregated evaluation measures while others investigate the consequences of evaluation-based incentive strategies on the teaching force.

We explore each of these dimensions in greater depth and contextualize the findings in several ways. First, we trace some of the historical factors and processes that led to IMPACT as it currently exists. Second, an extensive body of research has developed exploring issues related to the validity of measures of teacher effectiveness, the most comprehensive to date being the *Measures of Effective Teaching* project (Kane, Kerr, &Pianta, 2014). Almost all of these studies have been carried out as research projects independent of high-stakes evaluation systems. Nevertheless, we consider the IMPACT system in light of these research findings.

Third, we discuss IMPACT relative to other evaluation systems that are emerging across states. The IMPACT program has had a longer commitment than other systems to revamping its teacher evaluation and has also had more available financial and human resources as a result of funding from foundations and other organizations in the district's school reform initiatives (e.g., DC Public Education Fund). Also, DC is a single school district with one governing entity for non-charter schools. This stands in contrast with states (with the exception of Hawaii) that are

comprised of large numbers of local education agencies (i.e., school districts). Although 29 of the more than 50 public charter schools in the district participate in the teacher evaluation program, they are not subject to the same level of scrutiny nor have they had access to many of the infrastructure resources that have been available to the DC Public Schools (DCPS). Over the years, however, access to data and resource portals has increased for the participating charter schools (U.S. Department of Education, 2014a)

Methodology

This report is based primarily on a document review of publicly available information, most of which is accessible through the dcps.dc.gov website. Other sources for public documents were research studies, U.S. Department of Education websites and documents, and the website of the DC Public Education Fund (dceducationfund.org), which is an "independent not-for-profit formed to catalyze philanthropy in support of strategic reform in the DC Public Schools" (DC Public Education Fund, 2014a). In addition, we have been reviewing the evaluation systems of other Race to the Top states as part of another study and have included findings from this analysis as appropriate. We also requested additional information from DCPS during our conference calls, and they provided us with annual score distributions on component measures from the DCPS data warehouse. These component-level distributions across years have not been previously reported in any publicly available documents about IMPACT.

The quality of any teacher evaluation system needs to be considered in terms of a broad array of factors. Every system begins with one or more theories of action that describe how the process of scoring and the interpretation of scores used in evaluation will result in certain

outcomes and practices related to more effective teaching. In most models, including IMPACT, component scores are derived from the administration and application of a variety of measures that are then combined to yield aggregate scores that are used to support a range of decisions that include sanctions, incentives, and support.

We developed a conceptual framework to organize the information regarding state teacher evaluation systems that we were collecting. The framework includes major categories that describe not only the constituent measures but also the characteristics of how the overall system is designed, how it is implemented, what results are produced, what reports are disseminated, and what decisions are made and actions taken on the basis of these results. Categories relevant to IMPACT[1] include:

System Design – This category describes the overall evaluation system and includes:

a. major components of the evaluation;
b. performance levels for overall evaluations;
c. rules for including and excluding teachers from the evaluation;
d. designation of evaluators and how they are trained;
e. annual timeline for evaluations;
f. consequences associated with evaluation scores;
g. teacher recourse options;
h. prescribed growth plans for teachers based on ratings; and
i. data management system(s).

System Development – This category describes the processes, regulations, and other information related to the development of the system, including:

a. justification for the system;
b. legislation or administrative code that specifies system details;
c. stakeholder involvement, including any oversight evaluation committees;

[1] Several components from our framework are not included as they are not relevant to IMPACT (e.g., local flexibility within the system is not relevant because DC has only one district).

d. scope of any pilot/research studies;
e. major revisions to the system during development;
f. financial and non-financial support provided to the system;
g. training of leaders and teachers on components of the system, as well as the overall system; and
h. data and reports on evaluation outcomes (including reports on score distributions).

Student Growth Measures (SGMs) – This category describes the various measures intended to capture teacher effectiveness through the use of student achievement-related measures, including:

a. specification of measures to teachers depending on grade/subject assignment;
b. methods for aggregating growth measures into a composite growth score;
c. additional measurement components;
d. particular model used to estimate teacher contributions to student growth using standardized measures of achievement (e.g., value-added model [VAM], student growth percentile [SGP])
    i. the specific model and process/vendors used to generate the scores
    ii. performance levels derived from the model
    iii. subjects tested
    iv. teacher inclusion requirements
    v. individual, classroom, and other covariates
    vi. scoring processes (raw scores into SGM scores, treatment of multiple years);
e. quality control processes; and
f. assessment models that use student measures in non-tested subjects (Student Learning Objectives [SLOs], Teacher-Assessed Student Achievement Data [TAS])
    i. processes by which yearly SLOs are designed and developed
    ii. yearly requirements (e.g., number of SLOs, subjects)
    iii. evidence requirements and the process by which assessments used to score SLOs are chosen and approved
    iv. process by which assessments are scored
    v. quality control processes.

Teacher Practice Measures – This category describes the various measures associated with teacher practice, including:

a. construct focus;
b. instruments;

c. methods for aggregating different practice measures into a composite practice score;
d. specifics of the observation protocol
    i. domains assessed
    ii. observation cycle requirements
    iii. requirements for observers
    iv. scoring process;
e. professionalism; and
f. additional components and evidence collection.

Analysis and Reporting – This category describes how information in the system is disseminated and protected, including:

a. confidentiality of data;
b. reports to parents/public;
c. reports to teachers/districts;
d. reports to government agencies; and
e. reporting of measurement error.

Once the document review and analysis were completed, we developed a set of questions that was shared with a DCPS staff member who has responsibility for IMPACT. We then conducted a telephone conversation to review and get feedback on the contents of the analysis; to seek clarification on specific issues; and to obtain the most current updates regarding policy, practices, and documentation. As a result of these discussions, we were able to identify several documents that, though in the public domain, were not available at that time through the DCPS website.

Analysis

This section is divided into four sections. First, we describe the history and goals of IMPACT. Then, we review the design and implementation of its three major components: evaluation; decisions and actions; and feedback and support.

History and Goals

The vision of IMPACT began to take shape soon after Michelle Rhee took the position of Chancellor of DCPS schools in 2007. With the support of then Mayor Fenty, Rhee began work with her Chief of Human Capital, Jason Kamras, to research and develop a new teacher evaluation system (Curtis, 2011).

The project began with a year-long research and "listening" phase that included meetings with stakeholders. The second phase consisted of the creation of an IMPACT design team that included staff from the Human Capital division and individuals who were developing the district's Teaching and Learning Framework (TLF). The design team held focus groups and sought input from teachers and other staff to develop the IMPACT evaluation structure (Curtis, 2011). Since 2009, the fundamental goals of the system and the decisions that the evaluation is designed to support have remained essentially the same, although the components, weighting, and scoring methods have changed more substantially.

The underlying theory of action for IMPACT is explicit and presented on the DCPS website:

> *First introduced in 2009, the system is designed to help staff become more effective by:*

- ***Clarifying Expectations*** *– IMPACT outlines clear performance expectations that are tailored to staff members' specific job responsibilities.*
- ***Providing Feedback and Support*** *– Quality feedback is a key element to the improvement process, which is why IMPACT provides staff members with multiple opportunities to engage in conversations with their managers about strengths and areas for growth. IMPACT also provides data that helps instructional coaches, mentors, and other support personnel be more effective in their work.*
- ***Retaining Great People*** *– Having highly effective staff members in our schools helps everyone improve. IMPACT helps retain these individuals by providing significant recognition for outstanding performance.* (DCPS, 2014b)

The system is designed to support a set of critical decisions. First, evaluations are to be used to provide incentives and sanctions to teachers. The intent is to have an evaluation system that can support consequential human resource decisions, primarily around compensation and employment. Negative evaluations can lead to freezes in salary and employment separation (e.g., dismissal or transfer from the current school). High evaluation ratings can lead to substantial compensation incentives, especially for teachers in schools deemed "high poverty" and "low performing." Second, evaluations are intended to provide feedback and support to teachers, which assumes that the evaluative evidence is specific enough to support individual decision-making to improve practice. This second use relies not just on the assessment scores, but also on the clarification and understanding of the actions that led to the scores and support to address poorly assessed practices.

Design and Implementation of the Evaluation

IMPACT combines five annual measures into a single, overall summative score and focuses on student growth, classroom practice, practice outside the classroom, and

professionalism. The two measures of student growth are the Individual Value-Added (IVA) measure, a value-added model that looks at student growth across yearly administrations of standardized tests, and the Teacher-Assessed Student Achievement Data (TAS) measure in which the teacher and principal decide on growth goals given the specific class of students and the teacher must show evidence that students achieved those goals. Quality of classroom practice is measured through the Teaching and Learning Framework, a district-developed teacher practice rubric used to evaluate as many as five classroom observations during the school year. The Commitment to School Community (CSC) measure considers the teacher's performance outside the classroom—with colleagues, in support of school goals, and with parents. The final measure is Core Professionalism (CP), which functions somewhat differently than the other measures. When a teacher is identified as having acted unprofessionally, the CP measure leads to a reduction of his/her overall score, but the assumption is that teachers are professional in general; thus, the CP does not otherwise contribute to the aggregate score.

Scores on the IVA, TAS, TLF, and CSC measures are each transformed into scales that range from 1–4. The conceptual rationale for each measure, the implementation of each measure, and the derivation of scores for each measure are described below. We also report score distributions for each of the measures from the 2009–2010 school year to the 2012–2013 school year.

DCPS decided to launch a full-scale implementation and then make any necessary changes based on results rather than begin with smaller pilot implementations. Over time, there have been modifications in how specific measures are implemented and how they contribute to the final score. In addition, a measure of school value-added (Isenberg & Hock, 2012), which

was used in the earlier years of IMPACT, no longer is used to evaluate teachers. Specific changes are described in the relevant sections of this report.

Measures of Student Growth

Conceptual rationale. One of the most significant outcomes of the increased focus on teacher accountability has been the formal incorporation of measures of student growth into the regular evaluation of teachers. While researchers have experimented for decades with growth measures to compare classrooms, the push to incorporate them into formal evaluations with consequences began in the 1990s as researchers began testing value-added measures using statistical regression of standardized test scores to attempt to isolate a teacher's unique contribution to his/her students' learning (e.g., Wright, Sanders, & Horn, 1997). The theory underlying such a measure assumes that the teacher's net effect on learning can be calculated when external sources of variance in student growth are mathematically controlled, and that this effect can be meaningfully compared between teachers to infer relative effectiveness. While the specific external variables included in the regression differ from model to model, all value-added models control for each student's prior achievement over one or more previous years. The model may also include controls for student background variables (e.g., eligibility for free lunch, limited English proficiency, special education status) or classroom composition variables (e.g., percent of students in the class eligible for free lunch, average prior proficiency level) (e.g., Braun, 2005).

As value-added models began to be adopted in large-scale systems such as Tennessee (Wright et al., 1997) and educational economists and the accountability movement more generally continued to push for these measures to be used in high-stakes contexts (Goldhaber & Hansen, 2013; Hanushek, 2002; Kane, Taylor, Tyler, & Wooten, 2010), there was increasing public

attention and debate about the meaning of the measures and the validity of inferences that one may draw from different models. The considerable amount of estimated random error, the continued correlation of student background variables to teacher scores, the unobvious interpretation of the outcome score, and the argument that standardized test scores do not represent the important aspects of education have all raised strong criticisms in using VAMs in teacher evaluation (Baker et al, 2010; Haertel, 2013; Rothstein, 2010). Even those who consider VAMs more favorably raise cautions in their use (Ehlert, Koedel, Parsons, & Podgursky, 2012; Harris, 2011; McCaffrey, Lockwood, Koretz, & Hamilton, 2004).

As an alternative to value-added models, many evaluation systems are using student growth percentiles (SGPs) to accomplish the same goal (Betebenner, 2007). SGPs use quantile regression to calculate a percentile of growth compared to other students who scored at the same prior year's test score. While some models include some consideration for specific characteristics, SGPs tend not to control for the range of student background or classroom composition variables that value-added models do. Walsh and Isenberg (2013) compared the IVA and SGP models for the DC teachers and found that teachers who had larger proportions of English language learners and economically disadvantaged students would have somewhat lower teacher effectiveness scores using SGPs instead of VAM.

Both value-added models and student growth percentiles require multiple years of standardized test scores in order to calculate values that are comparable at the teacher and school levels. However, no Race to the Top system to date has standardized tests in the vast majority of subjects and grades, making it impossible to create meaningful growth scores from models that require common tests across students. Therefore, new teacher evaluations generally have an

additional measure of student growth, which is most frequently named a Student Learning Objective (SLO). SLOs may be based on commercially available assessments, developed individually by the teacher, or developed and set at the state, district, or school level.

While SLOs vary considerably in design and procedure from system to system, there are some shared characteristics. In the beginning of the year, the evaluator, often in collaboration with the teacher, sets one or more specific and measurable goals in light of baseline information about the teacher's students. Thus, as with other growth measures, there is an attempt to account for differences with respect to the prior academic achievement of the teacher's students. At the end of the year, after the teacher scores the student assessment(s), the evaluator assigns an evaluation score based on the extent to which the initial goals have been met.

While nearly every state requires SLOs for teachers in non-tested subjects, the general configuration of the measure differs from state to state. Some systems require SLOs for teachers in tested subjects as a measure of student growth in addition to the VAM or SGP in use. Some systems allow (or encourage) SLOs to target subgroups of students (such as low-achieving students or at-risk students) while others require SLOs to cover the entire class. In the design of the SLO, some systems provide complete flexibility for the teacher and evaluator to generate unique goals while others require SLOs to be comparable among teachers in the same grade and subject across a school, district, or the entire state. Even when such comparability is required, prior achievement of students is taken into account in setting the final target or evaluating the student performances.

IVA implementation. The IVA measure used in the IMPACT system is fairly typical of value-added models that include a sizable number of controls, as implemented in three other Race

to the Top states: New York, Louisiana, and Florida. Of the other Race to the Top states, three states use a value-added model with no student background or classroom composition control variables, eight states use SGPs with few or no background variables, and one state uses a substantially different model (the two remaining Race to the Top states did not have a model specified at the time of writing).

Critical to any teacher effectiveness measure based on student achievement scores are the policies and procedures for assigning and weighting student scores to an individual teacher. Comprehensive specifications for the procedures employed for estimating IVA in DCPS are detailed in a report by Mathematica Policy Research (Isenberg & Walsh, 2014), the contractor that performs this work for the district. We briefly overview critical aspects of the IVA design, but interested readers should refer to the Isenberg and Walsh study.

***Teacher inclusion*.** IVA estimates are available for only a subset of DCPS teachers: those who teach English language arts (ELA) for grades 4–10 or mathematics for grades 4–8. DCPS has stated for at least the last two yearly revisions of IMPACT that it will be able to calculate IVA estimates for a larger proportion of teachers by using additional standardized assessments for students in Kindergarten and first grade as well as those in high school English, mathematics, science, and social studies (DCPS, 2013a). There is no indication of when these new exams will actually be implemented.

Teachers with fewer than 15 students in either the current or previous year are excluded from the IVA measure. The 15 students can be distributed among multiple classrooms and grades, provided there are at least seven students per grade for a teacher in a given year.

Student inclusion. The IVA measurement model assumes that a teacher's contribution to

student learning is a linear function of the proportion of instructional time that the student was enrolled in class with that teacher. If all of the mathematics teaching[2] that a student experienced for a given year is provided by a single teacher, then the student's growth score is assigned fully to that teacher. If the student was only enrolled in that teacher's class for 50% of the time (e.g., the student moved from a different class or school mid-year), then the dosage, or relative weighting, of the student's growth score that contributes to the teacher's IVA estimate would only be 50%. Dosage is considered in terms of enrollment only, not attendance. Thus, a student who is enrolled for a full year in a given classroom but is absent for half of the classes would be assigned a dosage of 100%. Students are not included in a teacher's estimate if the dosage level is less than 5%. In cases of co-teaching in which both teachers are present in the classroom for the entire period, a given student contributes fully to each teacher's IVA estimates (Isenberg & Walsh, 2014).

Quality control of rostering. In order to accurately assign students to teachers, IMPACT has developed a yearly rostering process in which teachers indicate the subjects they teach and the students in those classes (DCPS, 2013a, pp. 11–13). This rostering is completed in the spring of each school year and then confirmed by school principals. Teachers may note circumstances such as the date the student transferred into or out of the class or if the student was pulled out of class for regular programs such as special education. Such factors are supposed to count toward the calculation of the proportion of instructional time attributed to the teacher. Instructions and an online training video are available to assist with the rostering process.

[2]Teaching contribution is defined by the formal class schedule. It is possible that a student learned some mathematics or language arts content from another teacher, of course, but unless designated as a mathematics or language arts class, respectively, the model ignores such potential influences.

IVA scoring. Scoring procedures are described fully in Isenberg & Walsh (2014).As noted, the IVA calculation weighs each student proportionally based on the percentage of instructional time with the teacher. Further, the teacher estimates are normalized based on the subject and grade. This minimizes the chance that systematic bias in the testing system or curriculum advantage or disadvantage teachers in specific grades but also ignores any real differences in overall teacher effectiveness that might exist among cohorts of teachers assigned to particular grades.

The statistical control variables that are included in the IVA estimation for each student are:

- pre-test in same subject as post-test;
- pre-test in other subject (e.g., control for mathematics while assessing reading growth);
- eligibility for free lunch;
- eligibility for reduced-price lunch;
- special education status;
- Limited English Proficiency (LEP) status; and
- attendance from the previous year.

In addition, the classroom-level control variables included are:

- class's average test score from the previous year;
- extent of the variation in the students' scores from the previous year; and
- proportion of students eligible for free or reduced-price lunch. (Isenberg & Walsh, 2014)

The model produces a percentile rank for each teacher's IVA estimate that is then transformed to a scale that ranges from 1.0 to 4.0. In the current model (Isenberg & Hock, 2012), teachers with an average IVA score (the 50^{th} percentile) receive a scaled score of 3.0, which is then used as input into the overall evaluation measure. This represents a change from the years of

IMPACT through 2010–2011 in which the average teacher received a scaled score of 2.5 (see The Education Consortium for Research and Evaluation [EdCORE], 2013, p. 43). While not explicitly discussed in the revisions of IMPACT, this may have been done to offset differences in the distributions of teachers with IVA compared to teachers without IVA. In the last year before the increase in the scale, both the average TLF and TAS scores were 3.0 (EdCORE, 2013), so teachers with IVA had a significant additional component that would have led to a decrease in their overall score compared to teachers in non-tested subjects.

IVA score distribution. Distributions of scores are not reported. However, in the current system, 50% of teachers receive an evaluation score of at least 3.0. Overall IVA ratings have increased due to the change in the scale between the 2010–2011 and 2011–2012 school years, but it appears from the technical documentation that the distributions would likely have remained stable since tg 2011-2012.

IVA rating quality. Mathematica, the district contractor, produces an annual final report that presents and describes the IVA model including covariates, statistical procedures, the characteristics of students and schools, and findings related to measurement error. These analyses help others interpret the quality and meaning of IVA ratings.

TAS implementation. The stated TAS requirements for teachers are that, "Assessments must be rigorous, aligned to the DCPS content standards, and approved by your school administration" (DCPS, 2013a, p. 42). Further, the district has published guiding materials to recommend assessments and goals for certain grades and subject areas (DCPS, 2011). Recommendations include the use of specific, commercially available assessments and standards-based assessments as well as suggestions for teachers to create assessments, projects, performance

assessments, and portfolios. Many suggestions for specific subjects and grades include multiple assessments that each target different instructional goals and are weighted to arrive at a final TAS score.

Nevertheless, these are guidance documents only, and the lone authorities to approve and oversee all aspects of the TAS measure are the evaluators, who are usually the school principal or assistant principal. While the district does review all TAS goals, it is only to "ensure they are workable" (DCPS, 2011, p. 2). The district does not provide examples of acceptable locally-developed TAS requirements and assessments, which contrasts with the kinds of supports for SLO measures that have been developed by other states.

The TAS measure is typical of SLOs in many states in which the teacher and evaluator have considerable flexibility in the design of the goals and the assessments. The TAS may involve one or multiple measures. The learning goals, assessments, scoring, relative weights (if multiple assessments are used), and evaluation criteria can all be negotiated between teacher and evaluator in the fall of the school year. IMPACT offers no explicit guidance or criteria for the approval of any of these TAS components. District material indicates that the teacher designs the measure and the administrator modifies or approves it (DCPS, 2013a), which is common in other Race to the Top systems. One of the only explicit restrictions from DCPS is that the TAS may not use the district-wide District of Columbia Comprehensive Assessment System (DC CAS) standardized test that is used to calculate the IVA as the underlying assessment so that evaluations of teachers of tested subjects are not overly reliant on a single test score.

We found no information about whether targets can be modified mid-year. In several other Race to the Top states, the assessment rubric is discussed during a mid-year conference and

potentially altered based on additional information on the students' starting point and level of progress at that point in the year.

TAS scoring. During the TAS development process, the administrator approves the scoring targets for the students and for the class as a whole, which should include a rubric through which the teacher will be scored at the end of the year. As with all IMPACT measures, scores range from 1 to 4; the performance levels for TAS are supposed to be defined based on assessments that indicate *little, some, significant,* and *exceptional* learning. Suggestions for the kinds of student achievement that would fall into different evaluative levels are shared with practitioners in the district: an example of *significant learning* (resulting in a 3.0) listed in the guidebook is, "1.25 years of growth" or "more than 80% mastery of standards," though it also states in a footnote that this is only "general guidance" (DCPS, 2013a, p. 46).

Teachers must present the evidence of the students' achievement to the administrator, and the administrator must verify the evidence and assign a score by the last day of school. If scores cannot be validated or the assessments used were not approved initially, the teacher receives a 1.

TAS score distribution. The score distributions for the TAS measure over the first four years of IMPACT are presented in Table 1.Numbers and proportion of teachers in each score range are reported. Several findings stand out. First, the majority of scores is greater than 3.0 (54% in 2009–2010 and 76% in 2012–2013). Second, scores have been increasing year to year. Third, the number of individuals with very low scores (less than 2.0) had hovered around 10% until 2012–2013, when it dropped to 6.4%.

TAS score quality. There is no systematic information collected as to the quality of TAS scores. There are no explicit standards of quality, no systematic mechanisms to review teachers'

scoring of student work or principal evaluation of the teacher's scoring. This lack of quality control of these locally developed measures is not unique to IMPACT. While this is typical in systems that rely on the teacher and principal to develop individual goals that need not be comparable across classrooms, several states that require specific assessments for their SLOs or student performances to be compared across grades and subjects have significantly more quality control. The quality of the TAS is not only unexamined but is almost totally dependent on the collective judgment and implementation of the teacher and administrator.

Table 1

TAS Score Distribution for School Years 2009–2010 to 2012–2013

	School Years							
Score Range	**2009–2010**		**2010–2011**		**2011–2012**		**2012–2013**	
	n	p	n	p	n	p	n	p
1.00–1.24	98	0.03	278	0.08	203	0.06	101	0.03
1.25–1.49	79	0.03	42	0.01	24	0.01	28	0.01
1.50–1.74	152	0.05	70	0.02	74	0.02	48	0.01
1.75–1.99	0	0	49	0.01	38	0.01	33	0.01
2.00–2.24	307	0.1	273	0.08	224	0.07	181	0.06
2.25–2.49	320	0.11	87	0.03	83	0.03	71	0.02
2.50–2.74	430	0.14	215	0.06	190	0.06	204	0.06
2.75–2.99	0	0	89	0.03	124	0.04	110	0.03
3.00–3.24	744	0.24	738	0.22	751	0.23	625	0.19
3.25–3.49	325	0.11	219	0.07	195	0.06	234	0.07
3.50–3.74	287	0.09	347	0.1	356	0.11	413	0.13
3.75–4.00	303	0.1	940	0.28	1,025	0.31	1,209	0.37
TOTAL	*3,045*		*3,347*		*3,287*		*3,257*	

Note: DCPS provided data; n = number of teachers; p = proportion of teachers in each score range

Measures of Teacher Practice: Teaching and Learning Framework (TLF)

TLF conceptual rationale. As with most teacher evaluation systems across the country, the classroom practice measure used in IMPACT is a significant contributor to the overall evaluation. Practice is most often evaluated through the use of an observation protocol. States and districts generally use either one of several commercially available protocols (most often the *Framework for Teaching* [Danielson, 2011]) or develop protocols of their own. For all of these instruments, the teacher is observed and rated on multiple dimensions of classroom practice several times during the school year. Each dimension is usually accompanied by descriptions of desired behaviors to support trained observers in reliably scoring teacher practice. Scores are typically aggregated across dimensions for a lesson, and lesson scores are averaged to derive an overall teacher practice score.

Research studies clarify that developing reliable and valid scores of teaching practice is challenging. Observers require substantial and ongoing training, and there is substantial variability in how raters assign scores and how scores and scoring vary over time and lessons (e.g., Casabianca, McCaffrey & Lockwood, 2014).Given such variation, research has documented the importance of reducing systematic error in estimating teacher practice quality by using multiple observations and multiple observers (Gitomer et al., 2014; Kane, Kerr, & Pianta, 2014; Mashburn, et al., 2014; Allen, Pianta, Gregor, Mikami, & Lun, 2011).

Construct focus and development. As described in the current guidebooks, DCPS (2013a) wanted a teacher practice framework that would be "a measure of instructional expertise" (p.6) and would reflect the "school system's definition of effective instruction, outlining key strategies which lead to increased student achievement" (p.12). The designed framework would:

- provide a common language for discussing teacher practice;
- allow for alignment of professional development (PD) to teachers' needs; and
- communicate clear performance expectations for DC teachers.

In order to develop the framework, DCPS, in conjunction with selected stakeholders, reviewed a large set of resources in 2008–2009, including teaching documents from states and professional organizations, observation protocols developed for research, teacher evaluation frameworks, and literature that presented research-based models for effective teaching (DCPS, 2013a). Following this review, a framework was developed to assess teaching practice over three domains: Plan, Teach, and Increase Effectiveness.

TLF implementation. While the Plan and Increase Effectiveness domains are described in the guidebooks, they have yet to be implemented. The Teach domain consists of nine standards that are scored holistically during observed lessons. Each dimension has its own rubric with descriptions provided at each of four performance levels: *Highly Effective, Effective, Minimally Effective*, and *Ineffective* (DCPS, 2013a).

In holistic scoring, the score is not to be based on a single incident during the observation, but rather on the overall impression of the observer for the standard. For 2013–2014, the standards are:

- lead well-organized, objective-driven lessons;
- explain content clearly;
- engage students at all learning levels in accessible and challenging work;
- provide students multiple ways to move toward mastery;
- check for student understanding;
- respond to student understanding;
- develop higher-level understanding through effective questioning;
- maximize instructional time; and
- build a supportive, learning-focused classroom community. (DCPS, 2013a, p. 15)

Annual cycle. DC IMPACT has developed business rules for the number and type of observations to be conducted each year. These rules are based on the level of the teacher in the career ladder that is described subsequently in this report. Most teachers receive four formal observations annually, but previously highly-rated teachers may receive as few as one formal observation per year. However, these teachers are permitted to request the full number of observations if they wish. A formal observation lasts at least 30 minutes, is unannounced, and is scored for all nine standards listed above. One additional, informal observation that does not count toward the overall score is also required to provide extra feedback to teachers.

Observations are conducted by school administrators, such as the principal or assistant principal, as well as by Master Educators. According to the guidebooks (e.g., DCPS, 2013a, p. 14), the Master Educator program designed by DCPS grew out of feedback from stakeholders that teachers would prefer to be evaluated by an impartial third party who had expertise in a relevant content areas well as experience in the classroom. DC IMPACT has recruited Master Educators from a nationwide pool, creating a current cadre across 13 content areas of 40 expert practitioners who not only conduct observations but also provide support for teachers in the district.

For a teacher receiving the full five observations (four formal and one informal), three are conducted by the school administrator and two by a Master Educator. As the teacher progresses along the Leadership Initiative for Teachers (LIFT) career ladder and the total number of observations is decreased, the number of observations by a Master Educator for a teacher at the Expert Teacher level eventually decreases to zero. For teachers in their first year, the first administrator observation is informal and announced, with the intent to provide useful feedback

before the formal scored observation cycle begins.

Throughout the process, feedback is supposed to be provided to the teacher. A conference within 15 days of an observation is required. For formal observations, this is followed by a full written report with scores and comments for each standard of the Teach domain. For teachers at the *Established* or above level, conferences only follow formal observations.

Master Educators participate in an extensive training program that includes a six-week summer institute in order to become certified (Curtis, 2011, p. 16) with ongoing training each summer (DCPS, 2014a). There is also a new online program called Align (DCPS, 2013d; DC Education Fund, 2014b) that was designed with a $1.5 million grant from the Bill and Melinda Gates Foundation and is available to all observers. This program consists of anchor videos for observers to rate in an attempt to increase evaluator reliability. If an observer's rating is not in line with the anchor score, the observer is provided with further training. While this program is available to all through its online platform, it is unclear whether there is a requirement for certified observers to continue to use Align and, if so, how often.

TLF scoring. During a formal observation, each dimension is scored 1–4, and the dimensions are then averaged for that observation. At the end of the cycle, formal observation scores are averaged to determine an overall practice score. DCPS is similar to other state systems in averaging multiple teacher observations in order to arrive at an aggregated teacher practice score, though there are certain alternative aggregation methods that assign overall scores based on the number of domains that are rated as *Effective* (or its equivalent) or higher rather than on an average score.

Beginning in 2011, the second year of implementation, a practice in which the lowest

observation score is dropped if it is more than 1 point below the average of all other scores (e.g., one observation score is 2 and the average of the others is 3 or higher) was adopted. The lowest score drop is available for all teacher career stages, and thus, if a teacher only has two observations, the lowest is still dropped if it is more than 1 point below the other. No other state appears to drop an outlying observation score.

TLF score distribution. The score distribution on the TLF measure over the four years for which data are available is presented in Table 2. A very small number of teachers are rated ineffective each year, and the majority of teachers receives scores in the *Effective* or *Highly Effective* range. The most recent year saw the highest proportion (69%) of teachers in that range.

TLF score quality. The use of Master Educators as outside observers with subject-matter expertise and independent from the school administrator/teacher relationship is unique among other Race to the Top evaluation systems. DCPS's choice of five observations for a summative score for early career teachers is higher than many other states' requirements and more in line with findings from research on the point at which observation scores converge (e.g., Bill and Melinda Gates Foundation, 2012). The use of multiple observers and multiple observations is in keeping with best practices identified in research (e.g., Bill and Melinda Gates Foundation, 2013). The provision of an informal and unscored observation for new teachers is found in many other states and is generally recognized as sound practice. The recent introduction of Align is a new innovation for sustaining and improving the expertise of observers in the system.

Table 2

TLF Score Distribution for School Years 2009–2010 to 2012–2013

	School Years							
Score Range	**2009–2010**		**2010–2011**		**2011–2012**		**2012–2013**	
	n	p	n	p	n	p	n	p
1.00–1.24	2	0	1	0	1	0	0	0
1.25–1.49	10	0	12	0	4	0	1	0
1.50–1.74	39	0.01	37	0.01	15	0	12	0
1.75–1.99	69	0.02	72	0.02	33	0.01	22	0.01
2.00–2.24	134	0.04	117	0.03	95	0.03	55	0.02
2.25–2.49	222	0.06	245	0.07	126	0.04	113	0.03
2.50–2.74	340	0.1	424	0.12	379	0.11	276	0.08
2.75–2.99	565	0.16	612	0.18	585	0.17	561	0.17
3.00–3.24	667	0.19	684	0.2	758	0.22	748	0.23
3.25–3.49	694	0.2	595	0.17	636	0.19	704	0.21
3.50–3.74	591	0.17	430	0.13	522	0.15	589	0.18
3.75–4.00	200	0.06	174	0.05	217	0.06	228	0.07

Note: DCPS provided data; n = number of teachers; p = proportion of teachers in each score range

Commitment to School Community (CSC) Measure

CSC conceptual rationale. Many education researchers have discussed the importance of

teachers' actions outside of planned lessons, particularly in how they interact with families, collaborate with teachers and support staff, and support school improvement efforts (e.g., Ladson-Billings, 2009). While classroom observation instruments and rubrics to assess lesson planning have been a research focus for decades, comparatively little research has looked into assessing and measuring teacher involvement outside instruction[3]. Similarly, inclusion of such measures in new teacher evaluations varies widely. For systems that use the full four-domain *Framework for Teaching*, Domain 4, *Professional Responsibilities*, contains one dimension for *Communicating With Families* and one dimension for *Participating in a Professional Community*. Other states define a broader rubric to capture some of these characteristics, often combining elements of professional behavior with those that demonstrate support for the school and community (e.g., Rhode Island Department of Education, 2014).

Washington DC's Commitment to School Community (CSC) measure is the only measure in a Race to the Top system that focuses entirely on extra-instructional participation as a separate entity from professional obligations. The stated intent of the measure is to reflect the extent to which the teacher supports and collaborates with the school community (DCPS, 2013a, p. 46). Teachers are assessed twice, the first before December 19 and the second before the end of the school year. The two scores are then averaged to arrive at the final score used in the evaluation rating.

CSC implementation. The CSC measure is composed of five different dimensions:

- Support for local school initiatives

[3]The National Board for Professional Teaching Standards (NBPTS) assessments included a measure that asked teachers to document their accomplishments outside of the classroom, but little research has been done on this type of measure.

- Support for special education and English language learner programs
- High expectations
- Partnership with families
- Instructional collaboration

CSC scoring. The school administrator conducts the CSC evaluation. The overall score for each of the twice-yearly assessments is the average of the dimension scores, and the final summative score is the average of the two assessments. This makes the final CSC score equivalent to a grand average of all dimensions across both assessment periods. All dimensions on the rubric are phrased in terms of the teacher acting "in an effective manner," and the difference between scoring levels is the frequency that these effective behaviors are observed. Teachers that *rarely* or *never* demonstrate the behavior receive a 1; *sometimes* equates to a 2; and *consistently* equates to a 3. For teachers to receive a 4, they must both consistently demonstrate the behavior as well as "extend impact" by independently and substantially contributing in additional ways (DCPS, 2013a, pp. 48–51).

CSC score distribution. The score distribution for the CSC measure across years is presented in Table 3. In 2009–2010, 74% of teachers received a score of 3 or above. Scores have increased to the point that 89% of teachers received a score of 3 or above by 2012–2013. While only 1.4% of teachers received a score less than 2 in 2009–2010, less than 0.4% of teachers (only 12 out of 3,294) received such a score in 2012–2013.

CSC score quality. We found no evidence of efforts to control the quality of CSC scores either through administrator training or implementation. The rubrics are written in high-inference language that is likely to be interpreted idiosyncratically by administrators in different schools, and the examples provided to guide assessment are brief and limited. No documentation exists to

clarify pivotal terms used in the rubric, such as *sometimes* or *effective manner*. It does not appear that there are any efforts to support the comparability of administrator scoring across the district.

Table 3

CSC Score Distribution for School Years 2009–2010 to 2012–2013

	School Years							
	2009–2010		**2010–2011**		**2011–2012**		**2012–2013**	
Score Range	n	p	n	p	n	p	n	p
1.00–1.24	6	0.00	1	0.00	3	0.00	1	0.00
1.25–1.49	8	0.00	5	0.00	2	0.00	1	0.00
1.50–1.74	12	0.00	13	0.00	2	0.00	5	0.00
1.75–1.99	22	0.01	13	0.00	16	0.00	5	0.00
2.00–2.24	68	0.02	40	0.01	36	0.01	23	0.01
2.25–2.49	99	0.03	63	0.02	37	0.01	37	0.01
2.50–2.74	236	0.07	199	0.06	129	0.04	119	0.04
2.75–2.99	496	0.14	333	0.10	188	0.06	179	0.05
3.00–3.24	1,248	0.36	1,102	0.33	931	0.28	751	0.23
3.25–3.49	552	0.16	532	0.16	546	0.16	507	0.15
3.50–3.74	478	0.14	576	0.17	703	0.21	781	0.24
3.75–4.00	284	0.08	459	0.14	758	0.23	885	0.27
TOTAL	*3,509*		*3,336*		*3,351*		*3,294*	

Note: DCPS provided data; n = number of teachers; p = proportion of teachers in each score range

Core Professionalism (CP)

CP conceptual rationale. In addition to measuring observable classroom practice, many states have made an effort to evaluate teacher professionalism. As discussed in the prior section,

the professionalism measure that arises in these other systems often includes components that are included in the CSC component of IMPACT. As with the observation framework, states typically either use an externally developed research instrument (such as Domain 4 of the Danielson *Framework For Teaching* or a subset of its dimensions) or develop a customized set of dimensions for the system. Race to the Top systems may assess professionalism as an individual component of teacher effectiveness or include dimensions of professionalism within a broader teacher practice measure.

The CP component of IMPACT is more limited and focuses on basic job responsibilities—coming to work on time, following policies and procedures, and interacting with people in a respectful manner. The district assumes that all employees will meet these employment obligations. Therefore, the CP measure is used only to deduct points from the overall evaluation for individuals who do not meet these expectations.

CP implementation. Professionalism has four components, referred to in the guidebook as "requirements for all personnel":

- Attendance
- On-time arrival
- Policies and procedures
- Respect

There are three levels of rating for each component: *meets standard*; *slightly below standard*; and *significantly below standard*.

CP scoring. CP components are rated twice annually, the first before December 19 and the second before June 19. Descriptions of each component and rating are provided on a rubric in the guidebook (DCPS, 2013a). In contrast to the CSC measure, the rubric descriptions are

comparatively low-inference and easily assessed statements. For example, to meet the standard for "On-time arrival," an individual must have "no unexcused late arrivals." To be classified as *significantly below standard* for "Respect," the teacher must demonstrate a pattern of failing to "interact with students, colleagues, parents/guardians, or community members in a respectful manner."

The Core Professionalism component only affects the teacher's score if there are professionalism "issues," and instead of contributing a weighted score to IMPACT's compensatory model, the CP score is disjunctive: if a teacher receives *slightly below standard* on *any* professionalism dimension for a half-year cycle, the overall CP rating for that cycle is *slightly below standard* and 10 points are deducted from the final summative score. If a teacher receives *significantly below standard* on *any* professionalism dimension for a half-year cycle, the overall rating is *significantly below standard* and 20 points are deducted from the final summative score. The same process is repeated at the end of the second cycle. Teachers shared between schools are assigned the lower of the two ratings, and the corresponding points are deducted from their final summative score. Therefore, teachers who receive the lowest rating of *significantly below standard* in both cycles will have their summative evaluation score reduced by 40 points.

CP score distribution. As seen in Table 4, most teachers' evaluation scores are not adversely affected by the CP measure, but deductions are not rare. As with the other measures, the overall scores for teachers have increased over time. During the first year of IMPACT (2009–2010), nearly one quarter of all teachers had some deduction, but less than 13% of teachers received a deduction in the 2011–2012 or 2012–2013 school years.

CP score quality. There is no evidence of quality control for the scoring of this measure

through training or during implementation. Scores are assigned at the discretion of the administrator.

Table 4

CP Proportion of Teachers with Score Deduction for School Years 2009–2010 to 2012–2013

	School Years			
Score Deduction Range	**2009–2010**	**2010–2011**	**2011–2012**	**2012–2013**
40-point Deduction	0	0.01	0.01	0.01
30-point Deduction	0	0.03	0.01	0.03
20-point Deduction	0.1	0.06	0.03	0.04
10-point Deduction	0.14	0.1	0.05	0.06
No Deduction	0.76	0.8	0.89	0.87

Note: DCPS provided data.

Aggregation of Measures

A critical design feature of any of the current evaluation systems is how components contribute to an overall evaluation metric. Many states use what is known as a conjunctive model that first transforms each summary score to a whole-number performance category and then uses decision rules to arrive at a final classification. For example, a state might say that if a score on student growth is "low," then it is not possible to have more than a mid-level final evaluation score.

IMPACT uses a compensatory algorithm that weighs and combines component scores to arrive at an overall, summative score for each teacher. Because all component measures are scored with a 1–4 range, the aggregation method is equivalent to a direct weighted average. Each component is multiplied by the weight of the measure, and the sum of all weights is 100. This leads to a summative score in the range of 100–400 points. The weights assigned to the two growth measures vary dependent on whether a teacher has an IVA score. DCPS refers to teachers

with an IVA score (i.e., teachers in tested subjects who have had the minimum number of students in their classroom) as "Group 1" teachers, and all other classroom teachers who do not have IVA scores that will be used in the final score calculation are "Group 2" teachers. The weights for each component measure are listed in Table 5 for Group 1 and Group 2 teachers.[4] For Group 2 teachers, the TLF score contributes 75% to the final evaluation score. As noted in the previous section, CP can lead to a deduction of points from this aggregated score.

Table 5

Current Weighting Scheme for IMPACT Components

	Group 1		Group 2	
Measure	**Weight**	**Score Range**	**Weight**	**Score Range**
IVA	35	35–140	-	-
TAS	15	15–60	15	15–60
TLF	40	40–160	75	75–300
CSC	10	10–40	10	10–40

Once the summative score is calculated, teachers are assigned to a final performance classification shown in Table 6. IMPACT classified teachers into one of four final performance levels prior to revisions made in the 2012–2013 school year. The original scale had levels (and ranges) of *Ineffective* (100–199), *Minimally Effective* (200–249), *Effective* (250–349), and *Highly Effective* (350–400). In 2012, a *Developing* level was added for scores of 250–299, which restricted the *Effective* level. Because the *Developing* level carries consequences for teachers who

[4]In June 2014, DCPS announced that it would not use IVA in teacher evaluation scores for the school year 2014–2015 only because of the introduction of the Common Core assessments. See http://www.washingtonpost.com/local/education/dc-public-schools-takes-hiatus-from-test-based-teacher-evaluations-as-city-moves-to-common-core-exams/2014/06/19/184b8b44-f7c2-11e3-8aa9-dad2ec039789_story.html

do not improve to *Effective,* this shifted the percentage from 50% to 66⅔% of possible points that a teacher must receive to be classified as performing satisfactorily.

Table 6

Performance Categories for IMPACT, Effective 2012–2013

Performance Level	Point Range	Percentage of Points Earned
Ineffective	100 and less than 200	0% – 33⅓%
Minimally Effective	200 and less than 250	33⅓% – 50%
Developing	250 and less than 300	50% – 66⅔ %
Effective	300 and less than 350	66⅔ % – 83⅓%
Highly Effective	350–400	83⅓% or more

The IMPACT system is relatively unique among Race to the Top evaluation systems in the configuration of its five performance levels. Most have four levels; for the few states that have five performance levels, none have three performance levels below *Effective* (or its equivalent). These systems retain two performance levels below *Effective* and then add two levels above *Effective*. As for the percentage of possible points that a teacher must receive to be classified as *Effective,* IMPACT falls well within the range of other states: Race to the Top evaluation systems range from 50% to 75% of possible points being required to achieve the lowest fully satisfactory performance level. While near 50% is the most common requirement, three of seven systems (including IMPACT) with compensatory aggregation approaches that set the percentages at the state level require that the teacher obtain 65% of the possible points in order to be considered *Effective*. The percentage of possible points is not a particularly informative

statistic on its own. These percentages must be considered in light of the distribution of individual component measures, discussed next.

Score distributions. Table 7 shows the distribution of overall evaluation scores across years. Year-to-year changes are difficult to interpret due to the changes that have occurred in the system. In 2009–2010 and 2010–2011, Group 1 teachers had IVA weights of 50% and did not receive TAS scores. Further, there was an additional aggregate score based on the IVA of all students in the school (School Value-added) that was removed. As noted in the section on IVA, scores were originally equated to a scale in which 2.5 was the IVA midpoint. This was changed to 3.0 in 2011–2012, possibly because all other component scores hovered around an average of 3.0 (with CSC still being somewhat higher). With an IVA midpoint of 2.5, teachers in tested subjects would be more likely to have a lower overall score simply due to the distribution of the components. This could result in unintended consequences such as providing a negative incentive to teach in the core subjects that had IVA scores available. Finally, the *Developing* performance level was introduced in 2012–2013.

Given the distribution of scores on the individual component measures, it is not surprising that most teachers in the district are rated *Effective* or *Highly Effective*. The inclusion of the *Developing* level, however, did seem to have a substantial effect on effectiveness ratings. Even while substantially fewer teachers received scores below 250 (i.e., they were rated in the *Ineffective* and *Minimally Effective* performance levels) in 2012–2013 than in earlier years, 2012-2013 also had the fewest number of teachers (75%) rated as at least *Effective*. A substantial number of those who would have been rated as *Effective* on the original four-level scale were now rated as *Developing.*

Table 7

IMPACT Teacher Effectiveness Performance Distribution for School Years 2009–2010 to 2012–2013

	School Years			
Performance Level	**2009–2010**	**2010–2011**	**2011–2012**	**2012–2013**
Ineffective	0.02	0.02	0.01	0.01
Minimally Effective	0.14	0.14	0.09	0.05
Developing				0.19
Effective	0.69	0.7	0.68	0.45
Highly Effective	0.16	0.14	0.22	0.3

Note: DCPS provided data.

Through their collective bargaining contract, teachers may appeal final summative ratings that are less than *Effective*, but only on procedural grounds. There must have been an error in terms of how the observations or assessments were performed or followed up in the conferences. The grievance must have the potential to lead to an improved overall evaluation classification in order to be considered. It must be filed within 14 days of the teacher's receipt of the score or it is considered "untimely."

To date, only a few states have reported the proportion of effective teachers in their teacher evaluation systems (U.S. Department of Education, 2014b), though that number should increase for school year 2013–2014. However, for the states that have reported these classifications, the findings are similar in that the majority of teachers is rated as *Effective* or better. The fact that most teachers are rated relatively highly is not at all surprising given the research on performance evaluation in general that shows that evaluations are most often highly skewed with relatively few low scores across different kinds of rating scales (e.g., Bretz,

Milkovich, & Read, 1992; Golman & Bhatia, 2012).

Design and Implementation of Decisions and Actions

The overall summative teacher score is used to support a number of critical employment decisions. Annual performance determines the extent to which teachers can advance on a professional career ladder with consequences for compensation and for reduced classroom observation requirements. In addition, teachers rated as *Highly Effective* have the opportunity to earn performance incentives. Finally, low performance ratings result in a range of employment sanctions.

Professional advancement. Teachers can ascend the DC IMPACT "Career Ladder," or LIFT (Leadership Initiative for Teachers) program based on a combination of experience and positive evaluation ratings, with specific requirements to advance levels. In order to advance, the ratings must be achieved consecutively. For example, the two *Highly Effective* ratings needed to advance from *Advanced* to *Distinguished* levels must be obtained in consecutive years. Movement occurs only in one direction—teachers can move up the ladder but do not move backwards if subsequent annual ratings are lower. Teachers also need to advance through each rung of the ladder. Two *Highly Effective* ratings are needed to become *Distinguished*, and then two additional *Highly Effective* ratings would be needed to become *Expert*. The levels and requirements are listed in Table 8.

Table 8

LIFT Career Ladder: Requirements for Advancement

Level	Requirements to Obtain
Teacher	None
Established Teacher	1 Highly Effective or 2 Effective
Advanced Teacher	1 Highly Effective or 2 Effective
Distinguished Teacher	2 Highly Effective
Expert Teacher	2 Highly Effective

Note: When multiple higher ratings are required to move levels, they must be in consecutive years.

According to the LIFT guidebook (DCPS, 2014d), as teachers move up the career ladder, they become eligible for additional leadership opportunities, including the ability to participate as curriculum writers, serve in policy fellowships, and help recruit and select new teachers for the school system. Teachers in advanced LIFT levels also qualify for a reduction in the number of formal classroom observations.

Compensation. DCPS negotiated with the Washington Teachers' Union (WTU) to substantially restructure the compensation structure. First, compensation is affected by a teacher's position on the LIFT career ladder, school assignment (high-poverty and low- performing), and whether one is a Group 1 (with IVA) or Group 2 teacher. Compensation comes in the form of base salary and single-year bonuses.

The LIFT compensation advances teachers on the district-negotiated salary guide in the

following ways:

- Advanced Teacher: +2 year service credit for base salary increases;
- Distinguished Teacher: +5 year further service credit, automatically moved to the master's degree base salary band if not there;
- Expert Teacher: +5 year further service credit, automatically moved to the PhD base salary band (DCPS, 2013b)

The second component of the compensation structure is the IMPACT*plus* program. Teachers who receive a final summative evaluation of *Highly Effective* qualify for annual bonuses that do not affect their base compensation. Annual bonuses range from $2,000 to $25,000 according to the following guidelines that take into account school free and reduced price lunch (FRPL), whether or not a teacher is in IMPACT Group 1, and whether or not the teacher is in one of the 40 lowest performing schools in the district. Teachers who are separated for disciplinary reasons, resign at the end of the school year, or are only part-time are not eligible for the bonus program. The bonus awards are specified in Table 9:

Table 9

Structure of Bonus Awards for IMPACTplus

	Teacher Group	
School Type	**Group 1**	**Group 2**
40 Lowest Performing	$25,000	$20,000
FRL Rate 60% or Higher	$15,000	$10,000
FRL Rate Less Than 59%	$3,000	$1,000

Note: DCPS provided data.

A key stipulation for receiving the bonus is that teachers must cede their contractual right to what is referred to as the “extra year” or other buyout options. District teachers who lose their teaching positions have the right to look for a new position for the next school year, with full

compensation and benefits. Teachers who are eligible for IMPACT*plus* bonuses must agree to waive this option in order to receive the additional compensation. While only about two-thirds of eligible teachers accepted the bonus during the first year of IMPACT, since that time, acceptance rates have been around 80%, as displayed in Table 10.

Table 10

DC IMPACT: Bonus Acceptance Rates

	School Years			
	2009–2010	**2010–2011**	**2011–2012**	**2012–2013**
% of Teachers Accepted	63.7	78.6	80.9	81.5

Note: DCPS provided data.

Funding for pay incentives. Much of the funding for the pay increase and bonus system comes from the work done by the DC Public Education Fund, a non-profit that solicits donations and grants in order to support quality teaching in DCPS and has contributed $12,420,367 to the DC IMPACT program in Academic Year 2011–2012 (From tax filing 2011, p.40).

Sanctions. There are several negative consequences attached to receiving a low overall evaluation rating. Teachers receiving any summative evaluation rating less than *Effective* have their salaries frozen by not advancing a step on the contractually-negotiated salary guide. Further, if teachers receive evaluation ratings of *Developing* for three years, they will also be subject to separation from their schools. For teachers who receive evaluation ratings of *Minimally Effective* for two years, they will also be subject to separation from their schools. Finally, teachers who receive a single evaluation rating of *Ineffective* will be subject to separation from their schools

immediately. Downward movement (*Developing* to *Minimally Effective*) results in the teacher being subject to separation from the school after the second year (i.e., the year in which he/she received the *Minimally Effective* evaluation). Upward movement (*Minimally Effective* to *Developing*) results in the teacher having a third year to improve before qualifying for separation. The principal, however, may recommend separation prior to these requirements if there is additional evidence that the teacher is not improving or if performance is declining and is already below *Effective*. Following the 2010–2011 school year, 98.5% of *Ineffective* teachers and 52.4% of *Minimally Effective* teachers left DCPS. Additionally, principals and instructional coaches are encouraged to prioritize teachers who score below *Effective* for professional development.

A recent study by Dee and Wyckoff (2013) was designed to evaluate whether teacher behavior was affected by the rewards and sanctions built into the system. Using a regression discontinuity design, they compared retention rates and subsequent IMPACT ratings by comparing teachers who were classified differently by the system yet had scores that did not differ from each other very substantially. Teachers could have overall scores that were similar, yet some would be labeled *Minimally Effective* and others *Effective*[5], while another group was on the cusp of *Effective* and *Highly Effective*. The most consistent finding of Dee and Wyckoff was that teachers who were labeled as *Minimally Effective* were more likely to exit the system than those who were labeled as *Effective.*

Design and Implementation of Feedback and Support

Overview of support. For teachers who are below *Effective*, the IMPACT guide states, "DCPS will encourage principals and instructional coaches to prioritize these teachers for

[5] The study was done at the time there were only four performance categories.

professional development in an effort to help them improve their skills and increase student achievement" (DCPS, 2013a, p. 64). Such guidance is less prescriptive than found in most Race to the Top states in which evaluation outcomes are more tightly tied to professional development and growth interventions. It is more typical for the feedback that accompanies evaluations to be required to be used in planning further professional improvement plans, and it is required that teachers rated unsatisfactorily be provided with additional resources and increased mentoring to address their assessed weaknesses. The IMPACT system leaves more room for judgment to the instructional coaches than in many other systems, and the relative effectiveness of this less prescriptive approach raises issues that can be explored empirically.

Instructional coaches. DCPS provides to every school instructional coaches whose jobs are specifically to provide support and feedback to teachers and the school leadership in a way that is comparatively uncommon in other systems (DCPS, 2014c). District leadership has stated that $15 million has been invested in instructional coaches (Sawyers, 2012). Coaches are tasked with analyzing data, designing professional development and support, and facilitating teacher learning. They are trained in the TLF and are informally encouraged to provide professional development surrounding the instrument's dimensions. Similarly, teachers are encouraged in the DCPS guidebook (DCPS, 2013a, p. 58) to share the results of their evaluations, but Master Educators are explicitly not allowed to share evaluation information with instructional coaches (Curtis, 2011, pp. 15–16). Per the WTU union contract, instructional coaches are forbidden from having evaluative duties or playing a role in the IMPACT evaluations (The Washington Teachers' Union, 2007). Curtis (2011) argues that this exclusion has led to a "firewall" between the instructional coaches and the Master Educators that prevents the teachers from receiving

differentiated support and increased the view that IMPACT is solely an accountability program.

Instructional coaches are paid on the general teacher salary schedule and are members of the Washington Teachers Union, must have at least three years of "successful" classroom teaching, and must be qualified for a teaching certificate in DCPS. Instead of teaching regular classrooms, however, they develop coaching plans to work with teachers and school leaders to facilitate understanding of new district initiatives (including IMPACT and the Common Core State Standards); conduct classroom observations and collect relevant artifacts to analyze teacher practice; and foster teachers' abilities to improve.

IMPACT evaluation of instructional coaches. Instructional coaches are evaluated in IMPACT using a separate instrument, the Instructional Coach (IC) Standards, that contains six dimensions (DCPS, 2013c). As with the teacher measures, there is a rubric that details performance based on a 1 to 4 scale. Each instructional coach is evaluated four times through the year, twice by a school administrator and twice by a member of the DCPS district office. The overall IC score is an average of all dimensions across all assessment periods. The IC score comprises 90% of the instructional coach's overall score and is combined with a 10% Commitment to School Community score and Core Professionalism score, each the same as with teachers.

Implementation, scoring, and score quality of instructional coach evaluations. The IC Rubric is similar to the CSC in that the details for each level are the same, but the difference in scores reflects increases in frequency. A score of 1.0 represents dimension behavior that is *rarely* or *never* observed, 2.0 represents behaviors that are *sometimes* observed, 3.0 represents *consistent* behavior, and 4.0 represents *consistent* behavior along with *extending impact.* Similar to our

commentary on the CSC measure, the rubric definitions are moderately high-inference statements with little explanation for administrators to evaluate what the pivotal terms mean.

If evaluation is supposed to intended to support professional growth, the explicit restriction of IC's not having access to evaluation data is potentially limiting, particularly given that the instructional coaches are charged with the following duties:

- Facilitating teachers' understanding and implementation of the Common Core State Standards and the DCPS Teaching and Learning Framework by developing and executing Collaborative and Individual Learning Cycles
- Creating detailed coaching plans, which include focused goals and measures of success, to drive learning cycles
- Supporting teachers' achievement of goals by using coaching strategies that gradually release responsibility for implementing instructional practices to the teacher (for example, co-planning, modeling, co-teaching, side-by-side coaching, and observing)
- Consistently analyzing teacher practice through ongoing classroom observations, data analysis, and examination of student work
- Providing clear and direct feedback to teachers based on analysis of practice
- Tracking student and teacher progress to assess the effectiveness of coaching
- Developing teachers' capacity to collect and analyze multiple sources of data to improve student learning
- Fostering collaboration and teacher leadership
- Participating actively on the school's Academic Leadership Team
- Attending professional development meetings, trainings, and all events led by the DCPS Instructional Coaching Program (DCPS 2013c)

Professional Development Resources

In addition to implementing IMPACT and the data management system to track

information, DCPS has developed several resources for teachers to engage in individual and independent professional learning and for instructional coaches to help teachers to meet their professional growth objectives. Most of these resources are accessed online, and there is little or no quality control to determine whether or not teachers are actually using them or whether or not they are successful in improving practice. While online resources solve accessibility and scheduling issues common for professional development programs, the lack of quality control surrounding their development and the lack of mechanisms to encourage their use may lead to underutilization of these tools:

- Professionally produced lesson videos from DCPS classrooms
- Curricular supports for the Common Core State Standards (CCSS)
- PD Planner—online catalog of professional development opportunities
- Educator Portal+—online platform to connect colleagues and resources
- Support for teachers focused on students with special needs, STEM, or International Baccalaureate
- The Washington Teachers' Union resources
  (DCPS, 2012xx)

Closing Comments

The teacher evaluation system in Washington, DC, has been at the forefront of a national movement to develop such systems. Given the attention IMPACT received as Race to the Top was designed, it is not surprising that it shares many features with other evaluation systems. However, it differs from other systems in important ways as well. We close by identifying several key common and differentiating characteristics of IMPACT.

First, the quality of an evaluation system ultimately comes down to the degree in which it leads to changes in teacher practice that increase student learning. It is inappropriate to think of these evaluation systems as grounded in traditional psychometric measurement practices that

focus on the internal validity and reliability of outcome scores. While value-added approaches to teacher evaluation have certainly been criticized by the measurement community (Baker et al., 2010), IVA is the only evaluation component in which the generation of scores can be understood in a way that meaningfully compares teachers or places them on an interpretable scale. TAS, TLF, and CSC scores are all produced with minimal evidence that scores are reliable or valid. As a general practice, recommendations that come from research regarding measurement in general and teacher evaluation measures in specific are largely unheeded.

That said, DC does have multiple observations and includes the Master Educators, independent observers who have subject-matter expertise and significant training and calibration that is largely in line with recommendations from research. The inclusion of judges of performance in the overall evaluation can provide an important perspective to the overall evaluation.

Second, IMPACT has maintained true to its initial vision of beginning with a full-scale implementation and then adjusting on the basis of data, including stakeholder reactions. IMPACT has made a substantial set of revisions and has been generally quite clear about the rationale for such changes.

Third, IMPACT has made a stronger commitment than any other current system to significant performance incentives. In most states, performance incentives are ultimately the province of local school districts. The consequences of these substantial performance incentives are now being studied and are likely to receive additional research scrutiny. Such studies may provide new insights on the extent to which large-scale evaluation-based performance incentives relate to desired outcomes.

Finally, IMPACT has made an extraordinary human and financial resource commitment to the entire teacher evaluation, performance incentive, and professional support efforts. In the current budgetary environment, this could only have happened with significant external resources. If DCPS IMPACT is successful in dramatically improving education, this infusion of external support must be given appropriate consideration in contemplating generalization of evaluation efforts to other locales.

References

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science, 333(6045),* 1034-1037. doi: 10.1126/science.1207998

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., . . . & Shepard, L. A. (2010). *Problems with the use of student test scores to evaluate teachers* (EPI Briefing Paper #278). Washington, DC: Economic Policy Institute. Retrieved from http://files.eric.ed.gov/fulltext/ED516803.pdf

Ballou, D., Sanders, W., & Wright, P. S. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*, *29*(1), 37–66. doi: 10.3102/10769986029001037

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment*, *17*(2–3), 62–87. doi: 10.1080/10627197.2012.715014

Betebenner, D. W. (2007). *Estimation of student growth percentiles for the Colorado Student Assessment Program*. Dover, NH: National Center for the Improvement of Educational Assessment.

Bill and Melinda Gates Foundation. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains* (MET Project Research Paper). Seattle, WA: Author. Retrieved from http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf

Bill and Melinda Gates Foundation. (2013). *Feedback for better teaching: Nine principles for*

using measures of effective teaching. Seattle, WA: Author. Retrieved from http://www.metproject.org/downloads/MET_Feedback%20for%20Better%20Teaching_Principles%20Paper.pdf

Braun, H. I. (2005). *Using student progress to evaluate teachers: A primer on value-added models*. Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/PICVAM.pdf

Bretz, R., Milkovich, G., & Read, W. (1992). The current state of performance appraisal research and practice: Concerns, directions, and implications. *Journal of Management, 18*(2), 321–352. doi: 10.1177/014920639201800206

Casabianca, J. M., Lockwood, J. R., &McCaffrey, D. F. (2014). Trends in classroom observation scores. *Educational and Psychological Measurement.* doi: 10.1177/0013164414539163

Casabianca, J. M., McCaffrey, D. F., Gitomer, D. H., Bell, C. A., & Hamre, B. K. (2013). Effect of observation mode on measures of secondary mathematics teaching. *Educational and Psychological Measurement, 73*(5), 757–783. doi: 10.1177/0013164413486987

Curtis, R. (2011). District of Columbia Public Schools: Defining instructional expectations and aligning accountability and support. Washington, DC: The Aspen Institute, Education & Society Program. Retrieved from http://www.aspendrl.org/portal/browse/DocumentDetail?documentId=1509&download

Danielson, C. (2011). *The Framework for Teaching evaluation instrument*. Princeton, NJ: The Danielson Group.

DC Public Education Fund. (2014a). *About us*. Washington, DC: Author. Retrieved from http://dceducationfund.org/aboutus.html

DC Public Education Fund. (2014b). Our work: Quality teachers and leaders. Washington, DC: Author. Retrieved from http://dceducationfund.org/our-work/excellent-teachers-and-leaders.html

District of Columbia Public Schools. (2011). *Teacher-Assessed Student Achievement Data (TAS) guidance*. Washington, DC: Author. Retrieved from http://tntp.org/assets/tools/DCPS_2011-2012TASGuidance_TSLT+3.12.pdf

District of Columbia Public Schools. (2013a). General education teachers with individual value-added student achievement data. Washington, DC: Author. Retrieved from http://dcps.dc.gov/DCPS/Files/downloads/In-the-Classroom/IMPACT%20Guidebooks/IMPACT-2013-Grp1-LR.pdf

District of Columbia Public Schools. (2013b). *IMPACT: The DCPS effectiveness assessment system for school-based personnel*. Washington, DC: Author. Retrieved from http://dcps.dc.gov/DCPS/Files/downloads/In-the-Classroom/Ensuring-Teacher-Success/2013-2014%20IMPACTplus%20For%20Teachers.pdf

District of Columbia Public Schools. (2013c). *IMPACT: The DCPS effectiveness assessment system for school-based personnel* (Instructional coaches). Washington, DC: Author. Retrieved from http://dcps.dc.gov/DCPS/Files/downloads/In-the-Classroom/IMPACT%20Guidebooks/IMPACT-2013-Grp15-LR.pdf

District of Columbia Public Schools. (2013d). *Rater training at DC Public Schools*. Washington, DC: Author. Retrieved from http://teachingislearning2013.org/sites/default/files/resources/G%20Rater%20Training%20and%20Certification.DCPS_.010312.v2.pdf

District of Columbia Public Schools. (2014a). *Additional support*. Washington, DC: Author. Retrieved from http://dcps.dc.gov/DCPS/About+DCPS/Career+Opportunities/Lead+Our+Schools/Master+Educators/Additional+Support

District of Columbia Public Schools. (2014b). *An overview of IMPACT*. Washington, DC: Author. Retrieved from http://dcps.dc.gov/DCPS/In+the+Classroom/Ensuring+Teacher+Success/IMPACT+(Performance+Assessment)/An+Overview+of+IMPACT

District of Columbia Public Schools. (2014c). *Instructional coaches*. Washington, DC: Author. Retrieved from http://dcps.dc.gov/DCPS/About+DCPS/Career+Opportunities/Teach+in+Our+Schools/Position+Overviews/Instructional+Coaches

District of Columbia Public Schools. (2014d). *LIFT: Leadership initiative for teachers, 2012–2013*. Washington, DC: Author. Retrieved from http://dcps.dc.gov/DCPS/Files/downloads/In-the-Classroom/IMPACT%20Guidebooks/LIFT-Guidebook.pdf

Dee, T., & Wyckoff, J. (2013). *Incentives, selection, and teacher performance: Evidence from IMPACT* (NBER Working Paper 19529). Cambridge, MA: National Bureau of Economic Research. Retrieved from http://cepa.stanford.edu/sites/default/files/w19529.pdf

The Education Consortium for Research and Evaluation. (2013). *Evaluation of the DC Public Education Reform Amendment Act (PERAA)*. Washington, DC: Author. Retrieved from http://dcauditor.org/sites/default/files/DCA132013.pdf

Ehlert, M., Koedel, C., Parsons, E. & Podgursky, M. (2012). *Selecting growth measures for school and teacher evaluations* (CALDER Working Paper 80). Washington, DC: National Center for the Analysis of Longitudinal Data in Education Research. Retrieved from http://files.eric.ed.gov/fulltext/ED535515.pdf

Gitomer, D. H., Bell, C. A., Qi, Y., McCaffrey, D. F., Hamre, B. K., & Pianta, R. C. (2014). The instructional challenge in improving teaching quality: Lessons from a classroom observation protocol. *Teachers College Record, 116*(6). Retrieved from http://www.tcrecord.org/Content.asp?ContentId=17460

Goldhaber, D., & Hansen, M. (2013). Is it just a bad class? Assessing the long-term stability of estimated teacher performance. *Economica, 80*(319). doi: 10.1111/ecca.12002

Golman, R., & Bhatia, S. (2012). Performance evaluation inflation and compression. *Accounting, Organizations, and Society, 37*, 534–543. doi: 10.1016/j.aos.2012.09.001

Grossman, P., Loeb, S., Cohen, J., Hammerness, K., Wyckoff, J., Boyd, D., & Lankford, H. (2010). *Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers' value-added scores* (NBER Working Paper No. 16015). Cambridge, MA: National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w16015.pdf?new_window=1

Haertel, E. H. (2013). *Reliability and validity of inferences about teachers based on student test scores* (William H. Angoff Memorial Lecture Series). Princeton, NJ: Educational Testing Service. Retrieved from https://www.ets.org/Media/Research/pdf/PICANG14.pdf

Harris, D. (2011). *Value-added methods in education: What every educator needs to know*. Cambridge, MA: Harvard Education Press.

Hanushek, E. A. (2002). Teacher quality. In L. T. Izumi & W. M. Evers (Eds.), *Teacher quality* (pp. 1–12). Stanford, CA: Hoover Institution Press.

Hazi, H. M, & Rucinski, D. A. (2009). Teacher evaluation as a policy target for improved student learning: A fifty-state review of statute and regulatory action since NCLB. *Education Policy Analysis Archives*, *17*(5), 1–22. doi: 10.14507/epaa.v17n5.2009

Isenberg, E., & Hock, H. (2012). *Measuring school and teacher value added in DC, 2011–2012 school year: Final report*. Washington, DC: Mathematica Policy Research. Retrieved from http://www.learndc.org/sites/default/files/resources/Measuring%20Value%20Added%20in%20DC%202011-2012.pdf

Isenberg, E., & Walsh, E. (2014). *Measuring teacher value added in DC, 2012–2013 school year: Final report*. Washington, DC: Mathematica Policy Research. Retrieved from http://dcps.dc.gov/DCPS/Files/downloads/In-the-Classroom/Ensuring-Teacher-Success/Measuring%20Value%20Added%20in%20DC%202012-2013.pdf

Kane, T. J., Kerr, K. A., & Pianta, R. C. (Eds.). (2014). *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching project*. San Francisco, CA: Jossey-Bass.

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2010). *Identifying effective classroom practices using student achievement data* (NBER Working Paper 15803). Cambridge, MA: National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w15803.pdf

Ladson-Billings, G. (2009). *The dreamkeepers: Successful teachers of African American children*.

San Francisco, CA: Jossey-Bass Publishers.

Mashburn, A., Downer, J., Rivers, S. E., Brackett, M. A.,& Martinez, A. (2014). Improving the power of an efficacy study of a social and emotional learning program: Application of generalizability theory to the measurement of classroom-level outcomes. *Prevention Science, 15*(2), 146–155. doi: 10.1007/s11121-012-0357-3

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: RAND Corporation. Retrieved from www.rand.org/content/dam/rand/pubs/monographs/2004/RAND_MG158.pdf

The New Teacher Project. (2011). *DCPS-TNTP IMPACT Conference: Overview and key lessons*. Brooklyn, NY: Author. Retrieved from http://tntp.org/assets/misc/20110603_IMPACT_conference_exec_summary_final.pdf?images/uploads/20110603_IMPACT_conference_exec_summary_final.pdf

Rhode Island Department of Education. (2014). *The Rhode Island model: Teacher evaluation and support system* (Edition II). Providence, RI: Author. Retrieved from http://www.ride.ri.gov/Portals/0/Uploads/Documents/Teachers-and-Administrators-Excellent-Educators/Educator-Evaluation/Education-Eval-Main-Page/Teacher-Model-GB-Edition-II-FINAL.pdf

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics*, *125*(1), 175–214. doi: 10.1162/qjec.2010.125.1.175

Sawyers, S. (2012, March). Q&A with Jason Kamras: What lies ahead for DC public schools. *The*

Hechinger Report. New York, NY: Teachers College at Columbia University. Retrieved from http://hechingerreport.org/content/qa-with-jason-kamras-what-lies-ahead-for-d-c-public-schools_8211/

U.S. Department of Education. (2014a). *Race to the Top District of Columbia report—Year 3: School year 2012–2013*. Washington, DC: Author. Retrieved from http://www2.ed.gov/programs/racetothetop/performance/dc-year-3.pdf

U.S. Department of Education. (2014b). *Race to the Top fund: Performance reports*. Washington, DC: Author. Retrieved from http://www2.ed.gov/programs/racetothetop/performance.html

Walsh, E., & Isenberg, E. (2013). *How does a value-added model compare to the Colorado growth model?* (Mathematica Working Paper). Washington, DC: Mathematica Policy Research. Retrieved from http://www.mathematica-mpr.com/~/media/publications/PDFs/education/value_added_colorado.pdf

The Washington Teachers' Union. (2007). *Collective bargaining agreement between the Washington Teachers' Union Local #6 of the American Federation of Teachers AFL-CIO and the District of Columbia Public Schools: 2007–2012* (Section 2.4.1.2.2). Washington, DC: Author. Retrieved from http://www.wtulocal6.org/usr/Documents/Final%20WTU%20DCPS%20Tentative%20Agreement.pdf

Wright, P., Horn, S., & Sanders, W. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education*, 11(1), 57–67. doi: 10.1023/A:1007999204543